Automatic Extraction of Standard Multimodal Knowledge for the Petroleum Field

Shan Huang*, Wangkai yue[a], Liuzhu jun[b], Wangyi fei[c], Yangxin xin[d]

Technology Inspection Center of Shengli Oilfield, SINOPEC, Dongying, Shandong, China

* Corresponding Author Email: huangshan.slyt@sinopec.com, [a]wangkaiyue.slyt@sinopec.com, [b]243766928@qq.com, [c]wangyifei524.slyt@sinopec.com, [d]yangxinxin.slyt@sinopec.com

Abstract. As the development of network and digitalization is accelerated, huge amount of multi-modal data is generated in the standardization activities of petroleum industry exploration and development, and the number of standards is large, covering a wide range of domains and being specialized. Based on the inherent characteristics of multi-modal data and standard knowledge organization framework, this paper proposes a method for automatic knowledge extraction of standard information from standard data of petroleum industry exploration and development. This paper advances the theory depth of knowledge organization of multi-modal standard knowledge and provides effective support for standard knowledge ordination, precise knowledge service and deep knowledge discovery in petroleum industry.

Key words: Automatic knowledge extraction, multimodal, exploration and development field, petroleum industry.

1. Introduction

With the digital transformation of the petroleum industry and the development of smart oil fields, a large amount of multimodal data such as text, images, videos, and audio has been generated through various standardization activities in the petroleum industry. However, the standard as an important means of optimizing industrial structure, promoting industrial development, and key link in the industrialization of innovative technologies is still in the early stage of digital transformation, causing a mismatch with the development of the oil industry [1]. Currently, standards are mainly in the form of paper and PDF, and the knowledge contained within these standards is obtained through manual reading, posing challenges for data sharing, information management, and data application mining. Traditional knowledge organization systems are mainly designed for single-modal data and cannot support the semantic representation, knowledge association, and integration of multimodal standard data in the petroleum industry. There is an urgent need for a more comprehensive knowledge extraction method to serialize and process the vast and diverse multimodal standard data in the industry.

2. General technical classification of standard knowledge in the petroleum industry

Standard literature in the petroleum industry has its own characteristics, and important knowledge information can be obtained from standard data catalogs, as well as from unstructured or semi-structured documents. The main function of text classification systems is to automatically categorize documents and assign them predefined category topics, facilitating the organization of electronic documents. Text classification in the petroleum industry mainly includes automated classification and rule-based classification technologies.

2.1. Automated classification technology

Typically, words are the most important classification knowledge. Based on the presence or absence, as well as the frequency of specific words in the standards, text categorization can be determined [2].For a specific classification problem, the words suitable for classification knowledge often only account for a small part of the dictionary, and these words also vary in their impact on classification.

In a classification problem distinguishing domestic and foreign news, country names, domestic and foreign place names, and personal names constitute the classification knowledge. Due to the limitations of existing word segmentation algorithms and dictionary size, many words or phrases with classification value in standard documents are not recognized, such as specialized terminology in various disciplines, as well as personal names, place names, and organizational names. To obtain more classification knowledge and improve the accuracy of classification, this paper adopts a statistical-based method for identifying compound phrases and undefined words, which can efficiently extract classification knowledge words or phrases from the text.

After obtaining more words to supplement the classification knowledge, it's necessary to determine which of these words are real classification knowledge and which are noise. The technical solution for this problem is feature extraction, which calculates the importance of each word for classification using statistical methods and selects the words with significant classification relevance, while removing unimportant or irrelevant feature words. The main algorithm for feature extraction is to construct a feature evaluation function, calculate the evaluation value for each feature independently, rank all features based on their evaluation values, and select the best features as the result's feature subset [3]. Common feature evaluation functions include mutual information between words and categories, information gain, expected cross-entropy, text evidence weight, odds ratio, and word frequency. Among them, expected cross-entropy (CHI) is one of the most effective methods.

The evaluation function only considers the frequency or distribution of feature words from a statistical perspective. It does not consider the actual distribution of feature words in specific documents. In fact, the classification impact of feature words in specific documents is influenced by factors such as document length, the number of similar feature words in the document, and the number of feature words in other categories. To compensate for this deficiency, this paper proposes a new feature extraction method based on the classification contribution of features, which improves the performance of feature extraction. After obtaining the classification knowledge, it's necessary to train a classifier to generate classification templates. The function of the classifier is to determine the category of a document based on the classification template. Effective classifiers include Bayes classifier, decision tree, KNN, and SVM.

Previous research on classifiers has often focused on improving the performance of individual classifiers. The main direction of this paper is to integrate these classifiers into a system to improve the overall classification performance of the system, and practical experience has proven this strategy to be very effective. SVM classifier and KNN classifier are currently the best classifiers. An ensemble classifier engine constructed by combining these two classifiers achieves better classification performance than the two individual classifiers mentioned above.

2.2. Rule-based Classification Technology

Based on rule-based automatic classification, users can randomly add or delete rules to meet personalized needs. The writing of rules follows logic operations such as (*) for "and," (+) for "or," (-) for "not," and (^) for "exclusive or."

Example: Originator= (Company A + Company B)-Main body=oil products or Title=oil products*Main body= (test method + test procedure). This allows for classification based on information such as the originator, title, and main body. Statistical algorithms are used for precise processing of classified text, rules, and categories.

To integrate the advantages of both automatic content-based classification and rule-based automatic classification, this paper combines the two forms of classification knowledge into one classification system, thereby improving the system's efficiency.

3. Technology for Text Clustering and Keyword Extraction in the Petroleum Field Standards (Title Data)

3.1. Standard Text Clustering Technology

The text clustering uses the K-means method, which is one of the most used clustering methods with good clustering results and fast speed [4]. The document similarity calculation method adopts the vector space model. Text feature selection includes words and n-grams to improve clustering performance. The system supports multi-level clustering and automatically generates multiple keywords for each category. The automatic clustering accuracy reaches over 75%, meeting the practical requirements of most applications. The clustering speed can reach 100 documents in 1 second and 10,000 documents in 5 minutes.

3.2. Standard (Title) Keyword Extraction Technology

Keyword extraction involves extracting words that are most relevant to the content of the standard text. Keywords are the words that best reflect the text's theme or meaning. The task of automatic keyword indexing needs to address two main problems:

How to extract candidate keywords from standard documents. This problem is fundamental, mainly involving word extraction. Generally, words are first identified using vocabulary-based methods, which do not require combination of string relationships during retrieval and thus have a fast retrieval speed. However, they have drawbacks such as construction difficulties and delayed updates. In addition, the number and quality of vocabulary entries directly affect the indexing quality and retrieval results. Another approach is statistical, using word extraction without a vocabulary list, or re-bundling fragments after segmentation, in which word matching and redundant filtering are crucial.

How to determine if a candidate is a keyword, and what is the basis for this. The first issue is crucial, mainly involving weight calculation. The extracted words are analyzed for their frequency in the text, relative word frequency, inverse document frequency factor, word position, part of speech, word value, word length, and introductions of statistical methods such as mutual information, TFIDF, maximum entropy, etc. The words are weighted accordingly, then sorted by weight, and some of the higher weighted indexing words are output.

This paper uses a combined rule-based and statistical keyword automatic indexing method, incorporating both domain keywords and automatically identified keywords using statistical methods, providing a more comprehensive approach to keyword extraction.

3.3. Synonym extraction technology in the petroleum field standard

The research on automatic synonym extraction technology can be mainly divided into aspects based on literal similarity, based on semantic dictionaries, based on co-occurrence relationships in large-scale corpora, and based on search logs. To improve the effectiveness of synonym recognition, this project uses a combination of multiple methods. If one word is a subclass of another, then there is a hierarchical relationship between them. For example, the relationship between "environmental protection" and "water environmental protection" is a hierarchical relationship. This relationship is asymmetric. We call the word with more specificity the subordinate term, and the word with more generality the superordinate term. Therefore, we can say that "water environmental protection" is a subordinate term of "environmental protection," and "environmental protection" is a superordinate term of "water environmental protection." To accurately determine the hierarchical relationship, this report conducts content theme analysis and indexing of standards in the petroleum field, analyzing and summarizing the standards of the same or similar directions or themes, and extracting relationships between different directions or themes based on this analysis.

4. Research on the extraction technology of specialized knowledge (metadata) in the petroleum field standards

4.1. Standard entity extraction technology

4.1.1 Extraction based on grammar

The extraction of standard entities (such as standardized objects) is mainly obtained from the title and scope of the standard. For the text of the title and scope of standard literature, noun phrases are extracted. When extracting and filtering, considerations are given to the parallel and modifying relationships between nouns, as well as the modifying relationship of adjectives to nouns. This ensures that the filtered part is a relatively complete ontology phrase.

When filtering according to grammar, it is necessary to analyze the grammar structure of the paragraphs, annotate the grammatical components in the paragraphs and their relationships. HanLP is used here as the syntax analyzer, as shown in Figure 1. HanLP is an NLP toolkit composed of a series of models and algorithms, led by big quick search and completely open source, aiming to popularize the application of natural language processing in production environments. HanLP has the characteristics of comprehensive functionality, high efficiency, clear architecture, up-to-date corpus, and customizable features.

```
private Set<String> targetFilter(String str) {
  List<String > target_list = new ArrayList<>();
  for (String s : SplitUtil.splitStr(str)) {
    Sentence sentence = this.segmenter.analyze(s);
    String last_word = "";
    String lastWordLabel = "w";
    String combo_word = "";
    for (IWord iWord : sentence.wordList) {
      String w = iWord.getValue();
      String flag = iWord.getLabel().toLowerCase();
      int length = target_list.size();
      if ("c".equals(lastWordLabel)){
        w = combo_word + "_" + w;
        combo_word = "";
      }
      if("c".equals(flag))
        combo_word = last_word;
      else if (flag.contains("n")) {
        if ("n".equals(lastWordLabel)){
          w = last_word + w;
          target_list.set(length - 1, w);
        } else
          target_list.add(w);
        flag = "n";
      }
      last_word = w;
      lastWordLabel = flag;
    }
  }

  Set<String> t_set = new HashSet<>(target_list.size());
  for (String target : target_list) {
    for (String t : target.split( regex: "_")) {
      t = StringUtils.strip(t);
      if (t.length() > 0 && t.length() <= 30 && SplitUtil.hasChinese(t))
        t_set.add(t);
    }
  }
  return t_set;
}
```

Figure 1. HanLP syntax analyzer

4.1.2 Filtering based on classifiers

Simply filtering the extraction indicators based on grammatical parts of speech does not yield accurate results as it contains many impure words. It is necessary to filter the candidate words again through a well-trained part-of-speech classifier. The extracted noun phrases are classified by the part-of-speech classifier. Phrases classified as ontology are then retained as the extraction results. On this basis, stop-word techniques are utilized for the following three functions: maintaining a stop-word table, ignoring meaningless ontology, ignoring English phrases, and ignoring excessively long phrases, as shown in Figure 2.

```
private Set<String> bertFilter(Set<String> words){
  Set<String> selectWords = new HashSet<>(words.size());
  if (words.isEmpty())
    return selectWords;
  ArrayList<String> wordList = new ArrayList<>(words);
  List<Pair<Double, CellType>> classifyResult = bertClassifyService.classifyForEach(wordList);
  for (int i = 0; i < classifyResult.size(); i++) {
    Pair<Double, CellType> typePair = classifyResult.get(i);
    if (typePair.getRight().equals(CellType.TI))
      selectWords.add(wordList.get(i));
  }
  return selectWords;
}
```

Figure 2. Stop-word techniques

4.2. Standard Attribute Extraction Technology

4.2.1 Design of the standard indicator attribute extraction model

The algorithm for extracting fragmented standard indicator attribute knowledge mainly includes table extraction and paragraph extraction. The implementation of these two modules depends on a part-of-speech classifier. Through table extraction, ontology and indicator items described in the table can be extracted. Through paragraph extraction, ontology described in the applicable scope paragraph of the standard literature can be extracted. Finally, the results of the two extraction parts are combined and added to the indicator database.

The flowchart of the standard indicator extraction algorithm is shown in Figure 3.

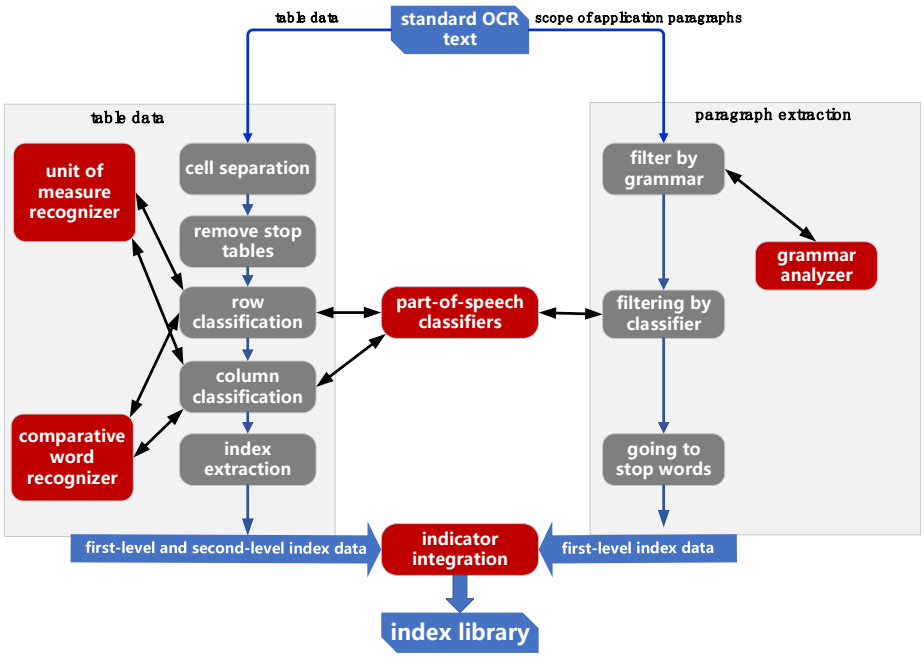


Figure 3. The flowchart of the standard indicator extraction algorithm

According to the flowchart of the standard indicator extraction algorithm in Figure 3, we conducted data training model testing, where the training data comes from the manually labeled data set in Oracle. We selected 800 standards from two professional fields in the upstream oil field for testing. Due to the relatively small amount of data in the limited range category, we used table names to add to the limited range organization to expand the data, making the data quantities of the four types consistent. Details can be seen in Table 1.

Table 1. Oracle manual labeling dataset

Table Head	Noumenon	Index item	Index value	Limit & scope
Oracle	1181	1672	2550	365

The mixed data of the four types are divided into ‘train’, ‘dev’, and ‘test’ sets in a 6:2:2 ratio for training and testing. The classification model first encodes the text using the BERT model. The BERT model, released by Google in October 2018, swept away the best results of 11 tasks in the field of NLP, making it the most important breakthrough in NLP today. The BERT model is obtained by training the Masked Language Model and predicting the next sentence task. This project uses a Chinese pre-trained model with 12 layers and 768 hidden units.

Next, the 768-dimensional vector output by the BERT model is passed through a fully connected network with a hidden layer and uses a dropout strategy. The output layer has 4 nodes, activated by the softmax function, and the 4-dimensional vector output represents the classification result, that is, the possibility under the four classifications. During the training process, cross-entropy is used as the loss function for the classification task, as shown in Figure 4.

```
def create_model(bert_config, is_training, input_ids, input_mask, segment_i
                 labels, num_labels, use_one_hot_embeddings):
  """Creates a classification model."""
  model = modeling.BertModel(
      config=bert_config,
      is_training=is_training,
      input_ids=input_ids,
      input_mask=input_mask,
      token_type_ids=segment_ids,
      use_one_hot_embeddings=use_one_hot_embeddings)

  output_layer = model.get_pooled_output()

  hidden_size = output_layer.shape[-1].value

  output_weights = tf.get_variable(
      "output_weights", [num_labels, hidden_size],
      initializer=tf.truncated_normal_initializer(stddev=0.02))

  output_bias = tf.get_variable(
      "output_bias", [num_labels], initializer=tf.zeros_initializer())

  with tf.variable_scope("loss"):
    if is_training:
      # I.e., 0.1 dropout
      output_layer = tf.nn.dropout(output_layer, keep_prob=0.9)

    logits = tf.matmul(output_layer, output_weights, transpose_b=True)
    logits = tf.nn.bias_add(logits, output_bias)
    probabilities = tf.nn.softmax(logits, axis=-1)
    log_probs = tf.nn.log_softmax(logits, axis=-1)

    one_hot_labels = tf.one_hot(labels, depth=num_labels, dtype=tf.float32)

    per_example_loss = -tf.reduce_sum(one_hot_labels * log_probs, axis=-1)
    loss = tf.reduce_mean(per_example_loss)

    return (loss, per_example_loss, logits, probabilities)
```

Figure 4. The flowchart of standard index extraction algorithm

The parameter settings during training are shown in Table 2.

Table 2. Parameter Settings

Parameter name	Parameter value	tips
train_batch_size	32	/
max_seq_length	128	Maximum sequence length
num_train_epochs	3	/
warmup_proportion	0.1	Preheating ratio of learning rate
learning_rate	5e-5	/

The final training result can reach 95% accuracy on the verification set.

4.2.2 Research on extracting indicator attributes from tables

Through extensive research on standard literature, it is found that the indicators in the standards of the petroleum industry mainly appear in the form of tables. Therefore, this paper focuses on the research of extracting indicator content from original standard tables. The process of extracting indicators from standard tables mainly includes the following aspects:

First, separate the table cells. The table with merged cells is separated into n independent cells, and the content of each cell remains the same as before merging, as shown in Figure 5.

OCR

item	Index mg/kg		
	pH<6.5	pH6.5—7.5	pH>7.5
HgT<	0.30	0.50	1.0
PbT<	40	30	25
CdT<	250	300	350
CrT<	0-30	0.30	0.60
/、/、/、^	150	200	250
DDD<	0.5	0.5	0.5

Separation

item	Index mg/kg	Index mg/kg	Index mg/kg
item	pH<6.5	pH6.5~7.5	pH>7.5
HgT<	0.30	0.50	1.0
PbT<	40	30	25
CdT<	250	300	350
CrT<	0-30	0.30	0.60
/、/、/、^	150	200	250
DDD<	0.5	0.5	0.5

Figure 5. Cell separation diagram

Second, generate a ternary data structure. An object with the following structure is generated for each cell in the table. Some attributes of the cell can be assigned different values based on the attributes of the row and column in which the cell is located, as shown in Figure 6.

```
class TableCell {
 String targetV1;
  String targetV2;
  String value;
  List<String> limit;
  String comparator;
  String unit;
}
```

Figure 6. Ternary data structure

Third, stop cells. Cells that are irrelevant to indicator extraction are ignored.

Fourth, column classification and row classification. Each row and column is traversed, and the content of each row (column) is treated as a set, and the entire set is input into a classifier to determine the type of the row (column). The corresponding type properties are assigned to the associated columns (rows), as shown in Figures 7 and 8.

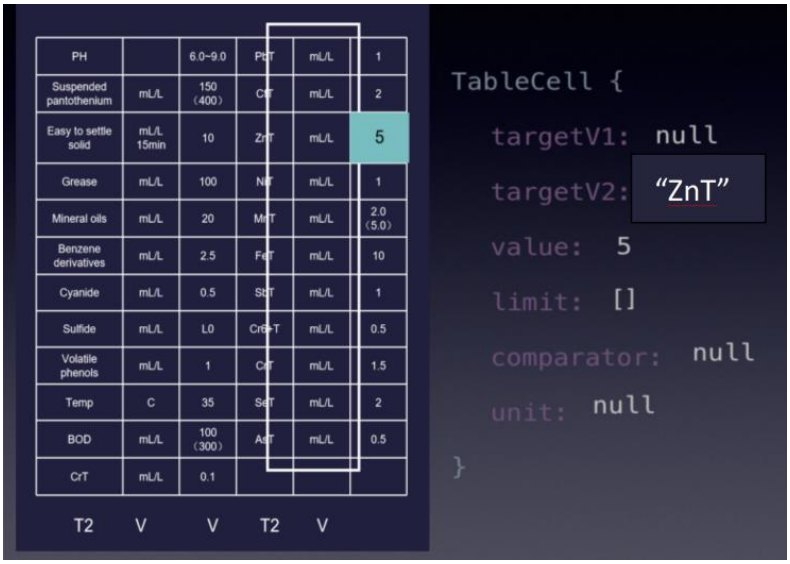


Figure 7. Column classification

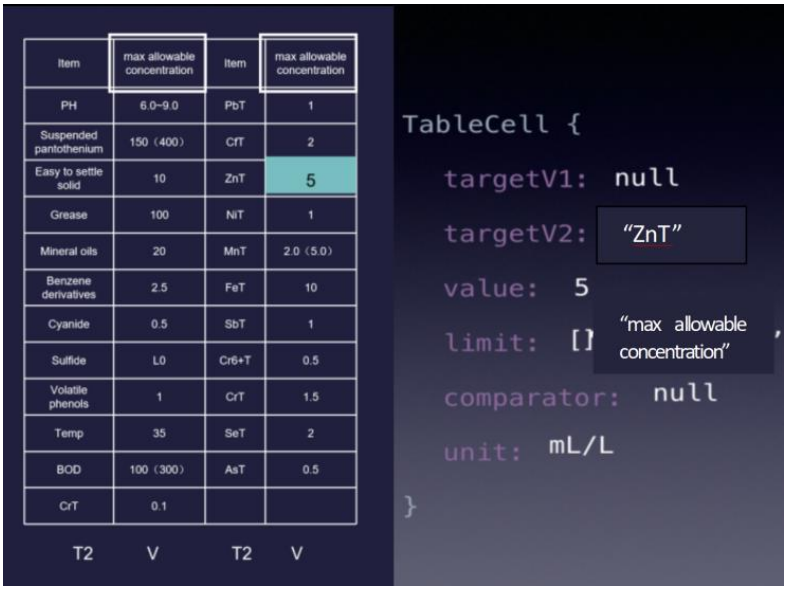


Figure 8. Row classification

When classifying rows and columns, it is also important to consider units of measurement and comparison words. The positions of units of measurement and comparison words in the table structure also show significant correlation. Classifying phrases containing units of measurement and comparison words into the part-of-speech classifier can lead to fluctuating results, resulting in incorrect classification (because units of measurement and comparison words are not included in the training data of the part-of-speech classifier). Therefore, before classifying rows and columns, it is necessary to first recognize, extract, and remove units of measurement and comparison words from the content of the cells to ensure that the data input into the part-of-speech classifier does not contain units of measurement and comparison words. Since the expression of units of measurement has strong regularity, the implementation of the unit of measurement recognizer is based on complex regular expression rules and considers the differences between Chinese and English expressions.

The comparator recognizer is implemented based on a self-built dictionary of common comparison words. The main comparison words include: “>“, “<“, “>=“, “<=“, “≥”, “≤”, “less than”, “greater than”, “greater than or equal to”, “less than or equal to”, “not greater than”, “not less than”, “not exceed”, “not surpass”, “not less than”, “not more than”, “not higher than”, “not lower than” and so on.

4.3. Relationship Data Extraction Technology

Entity relationship extraction commonly referred to as relationship classification, focuses on obtaining relationships, generating triplets in the form of <entity, attribute1, entity2>. Relationship extraction involves extracting various entity relationship information from texts and discovering relationships between entities to create an entity relationship network. Extraction techniques mainly include:

Rule-based extraction techniques, using manually created rules to handle information extraction issues in specific knowledge domains.

Statistical-based extraction techniques, which involve learning from annotated corpora, the system trained with rules can handle new texts that it has not seen before. The statistical method is mainly used to train the probability of a certain word being a part of a named entity based on a named entity corpus, and to use these probabilities to calculate the probability of a candidate term being a named entity, whereby terms with probabilities greater than a certain threshold are identified as named entities. The statistical-based method helps overcome the knowledge acquisition bottleneck in rule-based methods, and due to the higher degree of automation of tools used, this method is gaining increasing attention.

Combined statistical and rule-based extraction technique. Rule-based methods are subjective, have poor portability, and must deal with the inherent problem of ambiguity in language. Additionally, rules are difficult to cover all language phenomena, so it is hoped that machines will have learning ability in language processing. On the other hand, the application of human language is not purely a random process, and using purely statistical methods would result in a very large state search space. Leveraging rule knowledge and early pruning is a more effective method. The combination of rules and statistics can reduce the complexity and randomness of rule-based methods through probability calculations, as well as reducing the demands on the size of corpora for statistical methods.

5. Conclusion

With the development of information technology and the arrival of the era of big data, the development of petroleum informationization and digitalization puts forward new requirements for the management and utilization of standard data. Multi-modality complex data expressed in a form closer to human cognition world can better mine the complex deep-level relationships in the standard text of petroleum industry and realize the effective organization and value excavation of standard data resources. At present, the text semantic knowledge extraction is all based on “the paragraph” and “the word” analysis, which lacks cross-sectional semantic extraction, and petroleum industry, which involves a lot of fields, is a complex and extensive system, needs to consider the domain-specific

problems (different domains require customized development) and the data privacy problems (especially the relevant important contents such as technical, parameters and indicators in enterprise standards). Therefore, it is urgent to develop a more intelligent standard multi-modal knowledge extraction auxiliary tool based on the petroleum industry's business needs and scene, to provide auxiliary support for the construction of the big knowledge graph of petroleum industry in all domains, and it has important practical significance for improving the level of standard knowledge innovation services in the national energy industry.

References

[1] Junsheng Kong, Weizhao Li, Zeyi Liu, Ben Liao, Jiezhong Qiu, Chang-Yu Hsieh, Yi Cai*, Shengyu Zhang: Fast Extraction of Word Embedding from Q-contexts. CIKM 2021: 873-882.

[2] Liuwu Li, Yuqi Bu, Yi Cai*. Bottom-Up and Bidirectional Alignment for Referring Expression Comprehension. The 29th ACM International Conference on Multimedia (ACM MM 2021).

[3] Qi Peng, Changmeng Zheng, Yi Cai*, Tao Wang, Haoran Xie, Qing Li: Unsupervised cross-domain named entity recognition using entity-aware adversarial training. Neural Networks 138: 68-77 (2021).

[4] HUANG, Q. B., FU, M. Y., MO, L. Z, CAI, Y., XU, J. Y., LI, P. J., LI, Q. and LEUNG, H. F., 2021. Guided Question Generation with Contextual Structure and Sequence Information Capturing. In: Proceedings of the AAAI Conference on Artificial Intelligence, 35. 2–9 February 2021. AAAI Press.